IDA PAPER P-2998

A METHOD FOR SIMULATING CORRELATED RANDOM VARIABLES FROM PARTIALLY SPECIFIED DISTRIBUTIONS

Philip M. Lurie, *Project Leader*
Matthew S. Goldberg

October 1994

INSTITUTE FOR DEFENSE ANALYSES
1801 N. Beauregard Street, Alexandria, Virginia 22311-1772

IDA Log No. HQ 94-45764

DEFINITIONS

IDA publishes the following documents to report the results of its work.

Reports

Reports are the most authoritative and most carefully considered products IDA publishes. They normally embody results of major projects which (a) have a direct bearing on decisions affecting major programs, (b) address issues of significant concern to the Executive Branch, the Congress and/or the public, or (c) address issues that have significant economic implications. IDA Reports are reviewed by outside panels of experts to ensure their high quality and relevance to the problems studied, and they are released by the President of IDA.

Group Reports

Group Reports record the findings and results of IDA established working groups and panels composed of senior individuals addressing major issues which otherwise would be the subject of an IDA Report. IDA Group Reports are reviewed by the senior individuals responsible for the project and others as selected by IDA to ensure their high quality and relevance to the problems studied, and are released by the President of IDA.

Papers

Papers, also authoritative and carefully considered products of IDA, address studies that are narrower in scope than those covered in Reports. IDA Papers are reviewed to ensure that they meet the high standards expected of refereed papers in professional journals or formal Agency reports.

Documents

IDA Documents are used for the convenience of the sponsors or the analysts (a) to record substantive work done in quick reaction studies, (b) to record the proceedings of conferences and meetings, (c) to make available preliminary and tentative results of analyses, (d) to record data developed in the course of an investigation, or (e) to forward information that is essentially unanalyzed and unevaluated. The review of IDA Documents is suited to their content and intended use.

The work reported in this publication was conducted under IDA's Independent Research Program. Its publication does not imply endorsement by the Department of Defense, or any other Government agency, nor should the contents be construed as reflecting the official position of any Government agency.

IDA PAPER P-2998

A METHOD FOR SIMULATING CORRELATED RANDOM VARIABLES FROM PARTIALLY SPECIFIED DISTRIBUTIONS

Philip M. Lurie, *Project Leader*
Matthew S. Goldberg

October 1994

INSTITUTE FOR DEFENSE ANALYSES
IDA Independent Research Program

PREFACE

This paper was prepared by the Institute for Defense Analyses (IDA) under IDA's Independent Research Program. The objective was to develop an algorithm to simulate correlated random variables from partially specified, non-normal distributions.

This work was reviewed within IDA by Arthur Fries and Eleanor Schwartz.

CONTENTS

FIGURES

TABLES

I. INTRODUCTION

A. BACKGROUND

Cost estimates of major new programs in both government and industry have historically been unreliable. As a result, new programs frequently overrun their budgets. To remedy this situation, a formal risk analysis now accompanies most cost estimates. A risk analysis begins with an assessment of uncertainty in the individual factors or cost elements that contribute to cost growth. The risk analysis then aggregates these factors into a distribution of total cost. The decision-maker is thus presented with an entire range of total costs, along with associated probabilities, rather than a single cost estimate.

For some production processes, total cost may be expressed as the simple sum of the costs of lower-level components or cost elements. Using elementary statistical relationships, the first two moments of total cost can be obtained from the means, variances, and correlations among the individual cost elements. However, unless the individual cost elements are normally distributed, it is not generally possible to find a closed-form solution for the distribution of total cost. Therefore, it is not possible to go beyond moments and characterize the cost distribution in terms of percentiles, confidence bounds, or prediction limits.

In some cases, total cost may *not* be expressed as a simple sum, so even the moments of total cost are intractable. For example, alternative development programs may proceed in parallel, with the first successful arrival being adopted. Development cost in this case equals the full cost of the successful alternative (whose identity is not known in advance), plus the truncated costs of all unsuccessful alternatives. Thus, total cost becomes a complex, probabilistically-weighted average of truncated random variables.

Given such difficulties, simulation is often the only viable method for estimating the distribution of total cost. The input to a simulation must include a flow diagram indicating which processes proceed in parallel and which proceed in series. A distribution of cost must be associated with each arc (i.e., cost element) in the flow diagram. Typically, information on cost will come from many different subject matter experts. Usually, the only pieces of information these experts can provide are moments

(mean or median or mode, and possibly variance), marginal distributions (beta, triangular, log-normal, etc.), bounds (lowest and highest possible cost), if applicable, and correlations among the cost elements. Unfortunately, the correlation matrix can be specified independently of the marginal parameters for only a few multivariate distributions (the multivariate normal distribution, rarely used in cost analysis, is a notable example). In other words, once the marginal parameters have been determined, restrictions are placed on the correlations. These restrictions will often be incompatible with the correlation matrix specified by the user (or even estimated from data if the wrong distributional forms are assumed). Examples of restrictions on correlation matrices are provided in the next section.

B. RESTRICTIONS ON CORRELATION MATRICES

The following three examples illustrate the restrictions that certain multivariate distributions place on the correlation matrix.

1. Bivariate log-normal distribution [1]. Suppose Y_1 and Y_2 are jointly normally distributed with zero means, variances σ_i^2, $i = 1, 2$, and correlation ρ_Y. Let $X_i = \exp(Y_i)$ for $i = 1, 2$. Then X_1 and X_2 are jointly log-normally distributed with correlation:

$$\rho_X = \frac{e^{\rho_Y \sigma_1 \sigma_2} - 1}{\sqrt{e^{\sigma_1^2} - 1}\sqrt{e^{\sigma_2^2} - 1}}.$$

Thus, for example, if $\sigma_1 = 1$ and $\sigma_2 = 2$, the correlation between X_1 and X_2 must be within the bounds $-.09 \le \rho_X \le .67$. If the analyst believes that the correlation between X_1 and X_2 lies outside these bounds, then the bivariate log-normal distribution cannot be used.

2. Dirichlet distribution [1]. This is a type of multivariate beta distribution; its density is:

$$f(x_1,\ldots,x_n) = \frac{\Gamma(\sum_{i=0}^{n} \alpha_i)}{\prod_{i=0}^{n} \Gamma(\alpha_i)} \left(1 - \sum_{i=1}^{n} x_i\right)^{\alpha_0 - 1} \prod_{i=1}^{n} x_i^{\alpha_i - 1}.$$

The marginal distribution of any variable X_j is beta with parameters α_j and $\beta_j = \sum_{i=0}^{n} \alpha_i - \alpha_j$. The mean and variance of X_j are therefore $\alpha_j/(\alpha_j + \beta_j)$ and

$\alpha_j\beta_j/[(\alpha_j+\beta_j)^2(\alpha_j+\beta_j+1)]$, respectively. Furthermore, the correlation between any pair of variables X_i and X_j is:

$$\rho_{ij} = -\sqrt{\frac{\alpha_i\alpha_j}{\beta_i\beta_j}} \text{ for all } i \neq j.$$

Thus, once the means and variances of the individual cost elements have been estimated or specified, there are no degrees of freedom left to independently specify the correlation matrix.

3. Modified Farlie-Gumbel-Morgenstern (FGM) distributions [2]. This class of distributions was explicitly designed to allow flexible correlations with a given set of marginals. Consider univariate density functions $f_1(x_1)$ and $f_2(x_2)$ and let $\psi_i(t)$, i = 1, 2, be bounded nonconstant functions such that $\int_{-\infty}^{\infty}\psi_i(t)f_i(t)dt = 0$. Then the function defined by

$$h(x_1,x_2) = f_1(x_1)f_2(x_2)\{1+\omega\psi_1(x_1)\psi_2(x_2)\}$$

is a bivariate joint density with specified marginals $f_1(x_1)$ and $f_2(x_2)$, where ω is a real number satisfying the condition that $1+\omega\psi_1(x_1)\psi_2(x_2) \geq 0$ for all x_1 and x_2. Multivariate extensions of this family of distributions can also be derived [2].

If we let $\mu_i = \int_{-\infty}^{\infty} tf_i(t)dt$, $\sigma_i^2 = \int_{-\infty}^{\infty}(t-\mu_i)^2 f_i(t)dt$ and $\nu_i = \int_{-\infty}^{\infty} t\psi_i(t)f_i(t)dt$, and assume that all the integrals converge, we can show that the correlation between x_1 and x_2 is:

$$\rho = \frac{\omega\nu_1\nu_2}{\sigma_1\sigma_2}.$$

Suppose, for example, we wish to generate a modified FGM distribution with beta marginals on the interval (0,1) and parameters $\alpha_1 = 2$, $\beta_1 = 3$, $\alpha_2 = 5$, and $\beta_2 = 4$. To determine the range of possible correlations, we first approximated continuous functional forms for $\psi_1(t)$ and $\psi_2(t)$ by specifying $\psi_1(t)$ and $\psi_2(t)$ as constant within subintervals of length .05 (i.e., $\psi_1(t) = a_k$ and $\psi_2(t) = b_k$, $.05(k-1) < t < .05k$, for k = 1, ..., 20). Next, we determined the range of possible ω's that satisfy the constraint $1+\omega\psi_1(x_1)\psi_2(x_2) \geq 0$ for all x_1 and x_2. We then used a nonlinear optimization routine (Microsoft Excel Solver) to determine the values of a_k and b_k that maximize the range of possible correlations, subject to $\Sigma a_k = 0$ and $\Sigma b_k = 0$ (the latter constraints are necessary to ensure the uniqueness of a_k and b_k). The result was the following range: $-.65 < \rho < .65$. If the analyst believes that the correlation lies outside this range, then the FGM distribution cannot be used. Thus,

FGM distribution cannot be used. Thus, even though the FGM class was explicitly designed for flexibility, it still imposes serious restrictions on the range of possible correlations.

C. LITERATURE REVIEW

Many techniques are available for generating random numbers from univariate non-normal distributions; see Devroye [3] for a comprehensive survey. In the case of multivariate distributions, some techniques are available when particular distributions are fully specified (see Johnson [4]). However, our concern in this paper is with *partially specified* multivariate distributions. Some algorithms in the literature were derived assuming that the marginal distributions are fully specified, while other algorithms are based on the assumption that only the marginal moments (up to some order) are specified. The various algorithms also differ in the degree to which dependencies among variables are specified. Most algorithms require only the correlation matrix, but a few require higher-order product moments (i.e., moments of the form $E(X_i^p X_j^q)$ for positive integers p and q with $p+q>2$). For example, Parrish [5] presented a method for generating random variables from multivariate Pearson distributions, but his method requires knowledge of all product moments to fourth order.

Extending the ideas of Fleishman [6], Vale and Maurelli [7] developed a clever algorithm for generating correlated non-normal variables when the first four marginal moments (i.e., mean, variance, skewness, and kurtosis) are specified. They first generate correlated normal variables with an appropriate *intermediate* correlation matrix. They then apply cubic transformations to the normal variables, thereby fitting the four marginal moments. The cubic transformations convert the intermediate correlation matrix into the final, desired correlation matrix. They provide a system of non-linear equations to determine the intermediate correlation matrix from the desired correlation matrix. However, because polynomials in normal variables have infinite range, Vale and Maurelli's algorithm cannot be applied to distributions with bounded supports (e.g., the beta or triangular distributions often encountered in cost analysis).

Li and Hammond [8] developed an approach where the marginal distributions rather than the marginal moments are specified. Their approach is, in one sense, more general than Vale and Maurelli's, because the marginal distributions may have bounded supports. Once again, the major difficulty arises in determining the intermediate correlation matrix from the desired correlation matrix. Although Li and Hammond provide a formal solution to this problem, their solution entails inversion of double-

the integrand itself requires numerical approximation. As a practical matter, these integral equations may require a great deal of computation time if a high degree of accuracy is desired. In addition, because it treats each correlation independently, Li and Hammond's procedure will not necessarily yield an intermediate correlation matrix that is positive definite.

Lurie and Goldberg [9] developed an algorithm specifically designed to simulate random variables with bounded supports. Using a spectral decomposition, we first decompose the correlation matrix into its eigenvalues and eigenvectors. Because the eigenvectors are just linear functions of the original variables, their means, variances, and bounds are easily determined. Using these moments and bounds, a beta distribution is then fit to each eigenvector. Although the eigenvectors are uncorrelated by construction, they are not independent because domain restrictions on the original variables impose additional restrictions on the bounds of the eigenvectors. We found that a good approximation to the distributions of the domain-restricted beta variables could be obtained by using truncated beta distributions. The original random variables are then reconstructed as linear combinations of these truncated beta distributions.

The current paper improves upon our previous algorithm, in that the desired marginal distributions are reproduced exactly (within the sampling error of the underlying normal generator) rather than being approximated by linear combinations of truncated beta distributions. Moreover, the current method applies equally well to distributions with either bounded or infinite support, and even to "mixed" cases where some of the marginal distributions have bounded support and others have infinite support.

II. METHODOLOGY

A. THE PROPOSED SIMULATION ALGORITHM

A practical method for simulating correlated random variables should ideally have the following properties:

- the restrictions on the correlation matrix should be minimal, so it will usually be possible to generate a multivariate distribution compatible with the information the user has specified;
- the same method should apply to any arbitrary selection of marginal distributions;
- the simulated distribution should not depend on the order in which the variables are generated;
- the method should preserve the user-specified bounds, if any; and
- the method should be easily programmable on a computer and should not be so computationally intensive as to render it impractical.

The method proposed in this paper is an adaptation of an idea originally proposed by Li and Hammond [8]. Their procedure analytically relates the correlations of the desired variables to the correlations from a multivariate normal distribution. The desired correlations are obtained by transforming random variables from a multivariate normal distribution to a multivariate distribution with the desired marginals. This result is easily achieved by noting that if X has a standard normal distribution, then $\Phi(X)$ has a uniform distribution on the interval (0,1) and, consequently, $F^{-1}[\Phi(X)]$ has the desired distribution, where F is the desired cumulative distribution function (c.d.f.). If the desired output variables are standardized to have zero mean and unit variance, then the correlation between any pair of variables z_i and z_j can be written as:

$$\rho_{ij} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_i^{-1}[\Phi(x_i)]\, F_j^{-1}[\Phi(x_j)]\, \phi(x_i, x_j | r_{ij})\, dx_i\, dx_j\ ,\ 1 \le i < j \le k\ ,$$

where $\phi(x_i, x_j | r_{ij})$ is the bivariate normal density with correlation r_{ij}, and k is the number of variables.

Li and Hammond then propose to determine the input correlations r_{ij} that satisfy the equations:

$$\rho_{ij} - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_i^{-1}[\Phi(x_i)]\, F_j^{-1}[\Phi(x_j)]\, \phi(x_i, x_j | r_{ij})\, dx_i\, dx_j = 0\,, \tag{1}$$

for all $1 \le i < j \le k$. Although this method was proposed almost 20 years ago, it has never been widely used in practice because it is difficult to apply, and because the pairwise determination of normal correlation coefficients does not necessarily result in a positive semi-definite matrix.

The idea behind the approach taken in this paper is similar to the one just described, except that the multivariate normal correlations are determined empirically. The proposed procedure greatly simplifies the generation of correlated non-normal random variables. In fact, the examples presented later in the paper were calculated in a Microsoft Excel spreadsheet.

The steps in our approach are as follows:

Step 1. Generate n simulations of k independent unit-normal variables. Array these simulated values in an $n \times k$ matrix, denoted $\mathbf{X}$. This matrix remains fixed throughout the iterations.

Step 2. To obtain starting values, compute the Cholesky decomposition of the correlation matrix[1] ($\mathbf{R}_0$) desired for the final set of output variables; that is, determine the lower-triangular matrix $\mathbf{L}_0$ such that $\mathbf{R}_0 = \mathbf{L}_0\mathbf{L}_0^{\mathsf{T}}$.

Step 3. Linearly transform each *row* of independent normal random variables $\vec{x}_i = (x_{i1}, \ldots, x_{ik})$ into multivariate normal random variables $\vec{y}_i = (y_{i1}, \ldots, y_{ik})$ with correlation matrix $\mathbf{R}_0$. This is accomplished by the transformation $\mathbf{Y}_0 = \mathbf{X}\mathbf{L}_0^{\mathsf{T}}$, noting that if $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I})$, then $\mathbf{Y}_0 \sim N(\mathbf{0}, \mathbf{L}_0\mathbf{L}_0^{\mathsf{T}}) = N(\mathbf{0}, \mathbf{R}_0)$. The matrix $\mathbf{Y}_0$ has the same dimension as $\mathbf{X}$ ($n \times k$).

---

1 A correlation matrix is symmetric and positive semi-definite, or else a linear combination of variables could be found with negative "variance": $Var(\mathbf{c}^{\mathsf{T}}\mathbf{y}) = \mathbf{c}^{\mathsf{T}}\mathbf{R}\mathbf{c} \ge 0$ for all vectors $\mathbf{c} \ne 0$. In fact, a correlation matrix is positive definite unless a linear dependency exists (i.e., unless some linear combination of the variables has zero variance). Any symmetric, positive-definite matrix may be factored as $\mathbf{R} = \mathbf{L}\mathbf{L}^{\mathsf{T}}$ for $\mathbf{L}$ lower-triangular (Gill, Murray, and Wright [10], p. 36). The Cholesky decomposition is essentially the positive square root of a matrix. Finally, if the user-specified correlation matrix is not positive definite, a procedure described in a later section of this paper can be used to replace the user-specified matrix with a "nearby" matrix that is positive definite.

Step 4. Transform each element of $\mathbf{Y}_0$ using the unit-normal c.d.f., $u_{ij} = \Phi(y_{ij})$ for each $i = 1, \ldots, n$, and $j = 1, \ldots, k$. The columns of the $n \times k$ matrix $\mathbf{U}_0 = \{u_{ij}\}$ are uniformly distributed and correlated, but the correlation is no longer equal to $\mathbf{R}_0$ because the unit-normal transformation is non-linear.

Step 5. Invert the marginal uniform random variables generated in the previous step using the inverse c.d.f.'s of the desired marginal distributions. Let F_j denote the desired c.d.f. for the jth random variable. Then apply the transformation $v_{ij} = F_j^{-1}(u_{ij}) = F_j^{-1}[\Phi(y_{ij})]$ to the elements of the jth column of $\mathbf{U}_0$, $j = 1, \ldots, k$. The columns of the $n \times k$ matrix $\mathbf{V}_0 = \{v_{ij}\}$ will have the desired marginal distributions, but the correlations are further distorted from $\mathbf{R}_0$.

Step 6. Compute the empirical correlation matrix of the columns of $\mathbf{V}_0$. To do so, first standardize the columns of $\mathbf{V}_0$:

$$z_{ij} = \frac{v_{ij} - \bar{v}_j}{s_{v_j}},$$

where $\bar{v}_j$ and s_{v_j} are the empirical mean and standard deviation, respectively, of the jth column of $\mathbf{V}_0$, $j = 1, \ldots, k$. The empirical correlation matrix is computed as $\mathbf{R}_1 = \mathbf{Z}_0^{\mathsf{T}}\mathbf{Z}_0$, which is necessarily positive semi-definite and most likely positive definite.[2] For use in the next iteration, compute the Cholesky decomposition of the empirical correlation matrix, $\mathbf{R}_1 = \mathbf{L}_1\mathbf{L}_1^{\mathsf{T}}$.

Step 7. Evaluate a distance measure $\mathbf{D}$ between the empirical correlation matrix $\mathbf{R}_1$ and the target correlation matrix $\mathbf{R}_0$. Note that steps 3 through 7 define a mapping from a lower-triangular matrix $\mathbf{L}$ to a distance measure $\mathbf{D}$.

Step 8. Using a non-linear optimization routine, determine the elements of the lower-triangular matrix $\mathbf{L}$ that minimize the distance measure $\mathbf{D}$ computed in the previous step, subject to the constraints that $\ell_i \ell_i' = 1$ for all $i = 1, \ldots, k$, where ℓ_i is the ith row of $\mathbf{L}$. The constraints ensure that a valid correlation matrix is generated at each iteration.[3]

---

2 The empirical correlation matrix is at least positive semi-definite, because any quadratic form reduces to a sum-of-squares: $\mathbf{c}^{\mathsf{T}}\mathbf{R}_1\mathbf{c} = \mathbf{c}^{\mathsf{T}}\mathbf{Z}_0^{\mathsf{T}}\mathbf{Z}_0\mathbf{c} = (\mathbf{Z}_0\mathbf{c})^{\mathsf{T}}\mathbf{Z}_0\mathbf{c} = \sum_{j=1}^{n}(\mathbf{Z}_0\mathbf{c})_j^2 \geq 0$. This quantity can equal zero only in the unlikely event that a linear dependency exists among the columns of $\mathbf{V}_0$.

3 During step 3 of the pth iteration, columns of independent normal variables are transformed by the matrix $\mathbf{L}_p$ to induce the correlation matrix $\mathbf{R}_p = \mathbf{L}_p\mathbf{L}_p^{\mathsf{T}}$. This factorized representation of the correlation matrix is sufficient because a correlation matrix is symmetric and positive semi-definite, and any such matrix is subject to a Cholesky decomposition. Thus, we may restrict our search to lower-triangular matrices, for which roughly half the elements have known values (i.e., zero). Regarding constraints, Marsaglia and Olkin [11] have shown that, beyond symmetry and positive definiteness, the only additional constraints on a correlation matrix are unit diagonal elements.

Note that only one set of simulations is performed in this process—that of $n \times k$ independent unit-normal variables, which can be accomplished quickly and easily by means of the Box-Muller transformation (or by using whatever normal random number generator one has on the computer). At each iteration of the process, the transformations are updated and reapplied to the same simulated values. The random variables generated at the end of this process will have the desired marginal distributions and correlation matrix, provided the problem is feasible (i.e., provided a multivariate distribution with such a correlation structure can be reproduced via marginal transformations of a multivariate normal distribution). If the problem is not feasible, the procedure will yield a correlation matrix as "close" (in terms of the distance measure) as possible to the target matrix. In practice, most problems should be feasible unless the target correlation matrix is saturated with near-perfect dependencies.

Although the minimized value in step 8 of the algorithm is often known in advance to be zero, our interest centers instead on the lower-triangular matrix $\mathbf{L}$ at which the minimum occurs (i.e., the arg-min). This matrix is essential to the linear transformation in step 3 that induces the correct correlation matrix among the normal random variables. This situation is analogous to finding the root of an equation—the value of the function is known in advance, but we wish to locate the point at which that value is achieved.

The measure we use to determine the "distance" between the calculated and desired correlation matrices is the root mean square error (RMSE) between the corresponding matrix elements, i.e.,

$$RMSE = \sqrt{\frac{2\sum_{i=2}^{k}\sum_{j=1}^{i-1}(\rho_{ij} - \hat{\rho}_{ij})^2}{k(k-1)}},$$

where ρ_{ij} is the desired correlation between elements i and j, $\hat{\rho}_{ij}$ is the calculated correlation, and $k(k–1)/2$ is the number of unique off-diagonal elements in the correlation matrix.

In most applications this distance measure should be minimized at zero, that is, it should be possible to exactly achieve the desired correlation matrix. However, situations could conceivably occur where it may not be possible to achieve the desired correlation matrix by means of the proposed transformations of multivariate normal random variables. In these situations, the output matrix will be the best that can be achieved with the proposed methodology. In the many examples we tried with this method, the

transformations appeared to distort the original correlation matrix very little, so the desired correlation matrix was always achievable. This observation makes the target matrix a logical choice as a starting value for the original correlation matrix.

B. A MODIFIED SIMULATION ALGORITHM

Because our approach seeks to bring the correlations among the sample (simulated) random variables into conformity with the population (theoretical) correlations, the sample and population correlations will closely match, if not agree, regardless of the number of simulations performed. This property of our simulation method differs from standard simulation procedures, where sampling variability can result in the sample correlations being quite different from the population correlations.

The following example illustrates the importance of maintaining sampling variability in small samples. Suppose you have devised a procedure for testing the equality of parameters of several beta distributions. You are interested in the power of your testing procedure, particularly how power varies with the sample size. You expect your test to have low power to reject slight differences among beta distributions in small samples. You evaluate the power of the test through a Monte Carlo simulation, using the algorithm of this paper to generate beta variates with known parameters, then applying your test and attempting to discern the parameter differences. If our algorithm did not maintain sampling variability, you would reach far too optimistic an assessment of the power of your test in small samples.

The sampling variability in our procedure is reflected in the calculation of the lower-triangular matrix **L**, which will vary with each generation of the simulated independent normal random variables. That is, the matrix **L** is always adjusted to transform the new random variables so that the output correlation matrix is the one desired. Theoretically, the lower-triangular matrix **L** should depend solely on the population correlation matrix and be invariant with respect to the simulated independent normal random variables. If the matrix **L** is fixed, then the output correlation matrix will depend on the simulated random variables and may differ from the correlation matrix desired.

The theoretical equivalent to our method requires the determination of the lower-triangular matrix **L** such that

$$\sqrt{\frac{2\sum_{i=2}^{k}\sum_{j=1}^{i-1}(\rho_{ij}-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}F_i^{-1}[\Phi(x_i)]\,F_j^{-1}[\Phi(x_j)]\,\phi(x_i,x_j|r_{ij})\,dx_i\,dx_j)^2}{k(k-1)}}$$

is minimized, where $\{r_{ij}\} = \mathbf{R} = \mathbf{L}\mathbf{L}^{\mathsf{T}}$. When this minimization problem was programmed on a VAX 4000 Model 100 computer, however, one hour of central processing unit (CPU) time elapsed before completing even a single iteration. Thus, the theoretical determination of **L** is clearly impractical.

Although we cannot completely remove sampling variability from the empirical computation of **L**, we can modify our procedure by removing as much of the sampling variability as possible from the original, normally distributed observations. We will do this by adjusting the original observations so that they have exactly zero means, unit variances, and zero correlations.

First, the observations are standardized as

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_{x_j}},$$

where $\bar{x}_j$ and s_{x_j} are the empirical means and standard deviations of x_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, k$. Then the Cholesky decomposition $\mathbf{M}\mathbf{M}^{\mathsf{T}}$ of the sample correlation matrix **R** among the z's is computed and the observations are further transformed by

$$\mathbf{Y} = \mathbf{M}^{-1}\mathbf{Z} \tag{2}$$

to yield an array of uncorrelated observations. Applying our simulation algorithm to the observations generated in equation (2) will yield variables with the theoretical correlation structure. The algorithm will also determine the elements of the lower triangular matrix **L** needed to induce the appropriate correlations among the random variables generated in equation (2). Finally, steps 3 through 8 of the simulation algorithm are repeated to yield output random variables with sampling variation reflected in the computation of the sample correlation matrix.

C. GENERATING A POSITIVE SEMI-DEFINITE CORRELATION MATRIX

In many real-world situations, not enough data are available to accurately compute a correlation matrix. In this case, the correlations are frequently derived from expert opinion, often the combined opinions of many experts. The result is frequently an inconsistent correlation matrix, i.e., one that is not positive semi-definite. This inconsistency can also occur if the pairwise correlations are estimated empirically but are not all based on the same set of observations. The latter situation could arise if the number of observations on which all variables are observed is too small, and the analyst wishes to use all the available information to estimate each correlation.

To account for the possibility of an indefinite "correlation" matrix, we have developed a procedure for adjusting the user-specified matrix so that it is positive semi-definite and remains as "close" as possible to the original matrix. The procedure, which is similar to steps 3 through 8 of the simulation algorithm, determines the elements of a lower triangular matrix $\mathbf{L}$ that minimize the weighted quadratic distance between $\mathbf{LL}^T$ and the original correlation matrix $\mathbf{R}$. To ensure that $\mathbf{LL}^T$ is a correlation matrix, the diagonal elements of $\mathbf{LL}^T$ must be constrained to equal unity[4]:

$$\begin{aligned} \underset{\{l_{ij}\}}{\text{Minimize}} \quad & \left\|\mathbf{R}-\mathbf{LL}^T\right\| = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}(r_{ij}-s_{ij})^2 \\ \text{subject to:} \quad & s_{ij} = \sum_{h=1}^{k} l_{ih}l_{jh}, \quad i,j=1,\ldots,k \\ & s_{ii} = 1, \quad i=1,\ldots,k\,. \end{aligned} \tag{3}$$

In equation (3), r_{ij} are the elements of $\mathbf{R}$, l_{ij} are the elements of $\mathbf{L}$, and s_{ij} are the elements of $\mathbf{LL}^T$. The weights $\{w_{ij}\}$ are supplied by the user, and must be strictly positive to ensure a positive distance when $\mathbf{R}$ and $\mathbf{LL}^T$ are distinct. The user should supply a much larger weight to any particular correlations that are known with certainty (at least an order of magnitude larger). When a larger weight is applied, the minimization routine sets r_{ij} approximately equal to s_{ij} to avoid an otherwise large penalty.

To illustrate the matrix adjustment procedure, consider the following indefinite matrix:

---

4 As previously indicated, any correlation matrix may be factored as $\mathbf{LL}^T$; conversely, any matrix of the form $\mathbf{LL}^T$ may serve as a correlation matrix as long as the diagonal elements are constrained to equal unity. Expressing the correlation matrix in this form guarantees positive definiteness, and reduces the number of unknowns roughly in half compared to a dense (i.e., non-triangular) form.

$$R = \begin{bmatrix} 1 & .9 & .5 \\ .9 & 1 & .9 \\ .5 & .9 & 1 \end{bmatrix},$$

and suppose the weights for all elements are equal to 1.0. Based on the criterion in (3), the closest positive semi-definite correlation matrix is:

$$R_1 = \begin{bmatrix} 1 & .871 & .517 \\ .871 & 1 & .871 \\ .517 & .871 & 1 \end{bmatrix}.$$

Now suppose the analyst is very uncertain about the correlation between the first and third elements and specifies the weight for that correlation as .001, and weights for the remaining correlations as 1.0. The closest positive semi-definite correlation matrix is now:

$$R_2 = \begin{bmatrix} 1 & .9 & .619 \\ .9 & 1 & .9 \\ .619 & .9 & 1 \end{bmatrix}.$$

To the level of precision indicated, our procedure now places the entire adjustment on the relatively uncertain correlation between the first and third elements.

III. EXAMPLES

A. EXAMPLE 1

The first example considers the first-unit cost of a hypothetical 600-pound ultra-high frequency (UHF) satellite, consisting of ten components. Using "expert" opinion, we placed triangular distributions on each cost element to represent the risk associated with a hypothetical new satellite. The parameters of the triangular distributions are shown in Table 1.

Table 1. First-Unit Cost for 600-pound UHF Satellite

	Cost (Thousands of Dollars)				
Cost Element	Lower Bound	Mode	Upper Bound	Mean	Standard Deviation
Attitude Control	1,676	1,942	2,453	2,024	161
Electrical Power Supply	3,469	4,329	5,287	4,362	371
Telemetry, Tracking, and Command	860	1,014	1,671	1,182	176
Structure and Thermal	366	596	963	642	123
Apogee Kick Motor	201	314	402	306	41
Digital Electronics	5,433	8,431	8,828	7,564	758
Communications Payload	2,228	2,425	3,713	2,789	329
Integration and Assembly	544	691	1,011	749	97
Program Support	10,410	12,428	17,400	13,413	1,469
Launch Operations and Orbital Support	639	914	1,030	861	82

We estimated the target correlation matrix from historical data, but data were missing for some of the costs elements of some of the historical systems. In an effort to maximize sample size, we estimated each individual correlation from the largest possible subsample containing data on the corresponding pair of cost elements. Thus, the exact subsamples varied from one correlation to another. The resulting correlation matrix $\{\rho_{ij}^0\}$ was indefinite, containing one negative eigenvalue and all others positive. We applied the procedure described in the previous chapter to replace this indefinite matrix with a "nearby" positive semi-definite matrix, $\{\rho_{ij}^1\}$. Our procedure adjusted all of the eigenvalues and, in particular, replaced the single negative eigenvalue with the value of zero, resulting in an adjusted correlation matrix that was positive *semi*-definite (but not positive definite). This result occurred because our procedure attempts to *minimize* the distance between the adjusted and

unadjusted correlation matrices. Any larger adjustment, yielding a strictly positive eigenvalue, would have increased the value of the distance measure.

Using Li and Hammond's procedure, we solved equation (1) to determine the correlations among normal variables necessary to achieve the adjusted correlations among the transformed variables.[5] Li and Hammond's procedure adjusts each correlation *individually*, mapping the adjusted target correlations $\{\rho_{ij}^{1}\}$ into the corresponding normal correlations $\{r_{ij}\}$. We have observed empirically that the normal correlations $\{r_{ij}\}$ differ only slightly from the adjusted target correlations $\{\rho_{ij}^{1}\}$. In this particular instance, $\{r_{ij}\}$ differed from $\{\rho_{ij}^{1}\}$ in that the zero eigenvalue was replaced by a negative value. Thus, the normal correlation matrix resulting from Li and Hammond's procedure was indefinite (i.e., logically inconsistent).[6]

Although this example may appear pathological, we believe that it actually represents a common situation in cost analysis. Correlations among cost elements are typically estimated either from unbalanced data sets (as in the example above), or else from expert opinion. In either case, the target correlation matrix is likely to be indefinite, so that the adjusted target matrix is only positive *semi*-definite. Li and Hammond's solution for the normal correlations is more likely to be indefinite when the adjusted target matrix is positive semi-definite than would be the case if the eigenvalues were strictly bounded away from zero.

Next, we used our own algorithm to generate 1,000 simulations of the desired multivariate distribution, using the procedure outlined in steps 1 through 8 in the previous chapter. Our simulation algorithm resulted in a positive-definite solution because the sample, rather than the population, correlations were estimated. Thus, the correlations among the simulated values were identical to the population correlations (i.e., $RMSE = 0$).

It remains to show that the simulated marginal distributions are consistent with triangular distributions having parameters given in Table 1. Plots of the empirical distribution function for each simulated variable against its theoretical counterpart revealed a close fit in each case. The worst-fitting case (the case with the largest maximum difference between the theoretical and empirical distributions), corresponding to Program Support Costs, is shown in Figure 1.

---

5 For each pair of variables, the solution to equation (1) was determined using International Mathematical and Statistical Language (IMSL) subroutines DTWODQ to perform the double integration and DZREAL to find the roots. All calculations were performed using double-precision arithmetic.

6 Li and Hammond ([8], p. 561) were well aware of this possibility.

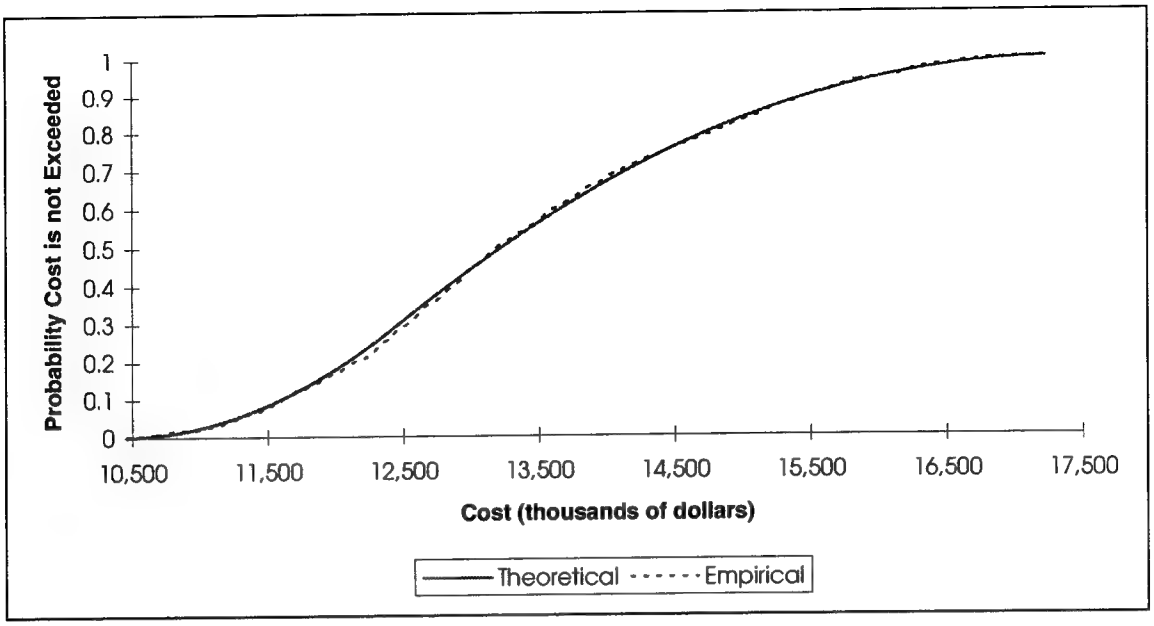

Figure 1. Theoretical and Empirical Distributions of Program Support Cost (1,000 Simulations)

The fact that, in the worst case, the two distributions are almost indistinguishable implies that the fits are close in every case. Accordingly, the summary statistics associated with each distribution also correspond closely. Comparisons of the true population parameters with those estimated from the simulation are shown in Table 2. It is clear from Table 2 that the simulation has done a good job of reproducing the population parameters. Furthermore, the distributions are all approximately triangular and have the desired correlations (to two decimal places).

Table 2. Comparison of True and Estimated Parameters From Example 1

	Mean		Standard Deviation	
Cost Element	True	Estimated	True	Estimated
Attitude Control	2,024	2,025	161	162
Electrical Power Supply	4,362	4,365	371	374
Telemetry, Tracking, and Command	1,182	1,179	176	177
Structure and Thermal	642	644	123	125
Apogee Kick Motor	306	305	41	41
Digital Electronics	7,564	7,554	758	770
Communications Payload	2,789	2,788	329	321
Integration and Assembly	749	748	97	95
Program Support	13,413	13,424	1,469	1,453
Launch Operations and Orbital Support	861	861	82	83

B. EXAMPLE 2

The second example uses the same data as the first, but we generated only 100 simulations and we employed the modified simulation algorithm (see Section II.B) to take account of sampling variability. Comparisons of the true population parameters with those estimated from the simulation are shown in Table 3. A plot of the worst-fitting distribution (corresponding to Program Support Costs) is shown in Figure 2, and a comparison of the population correlations with the sample correlations is displayed in Figure 3. The 45°–line in Figure 3 represents the situation where the population and sample correlations are equal. When applied to the adjusted normal random variables (i.e., adjusted to have zero means, unit variances, and zero correlations) the *RMSE* was 0 (to two decimal places). However, when applied to the original normal random variables, the *RMSE* increased to 0.08.

Table 3. Comparison of True and Estimated Parameters From Example 2

Cost Element	Mean		Standard Deviation	
	True	Estimated	True	Estimated
Attitude Control	2,024	2,014	161	163
Electrical Power Supply	4,362	4,308	371	388
Telemetry, Tracking, and Command	1,182	1,154	176	166
Structure and Thermal	642	627	123	121
Apogee Kick Motor	306	296	41	45
Digital Electronics	7,564	7,344	758	872
Communications Payload	2,789	2,753	329	310
Integration and Assembly	749	736	97	97
Program Support	13,413	13,372	1,469	1,427
Launch Operations and Orbital Support	861	852	82	80

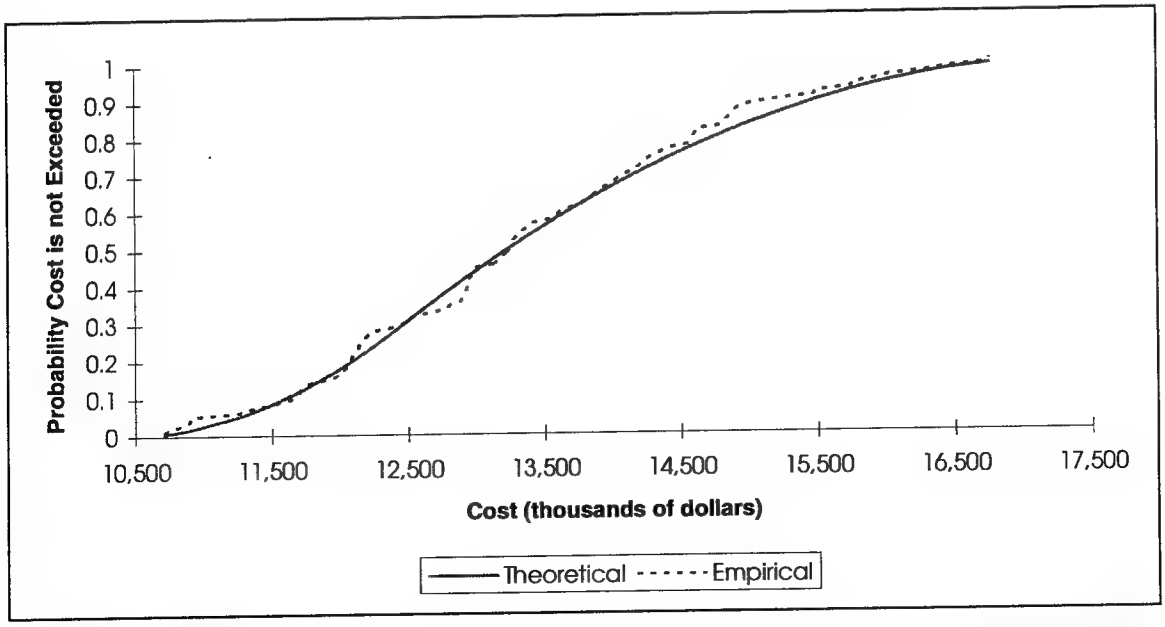

Figure 2. Theoretical and Empirical Distributions of Program Support Cost (100 Simulations)

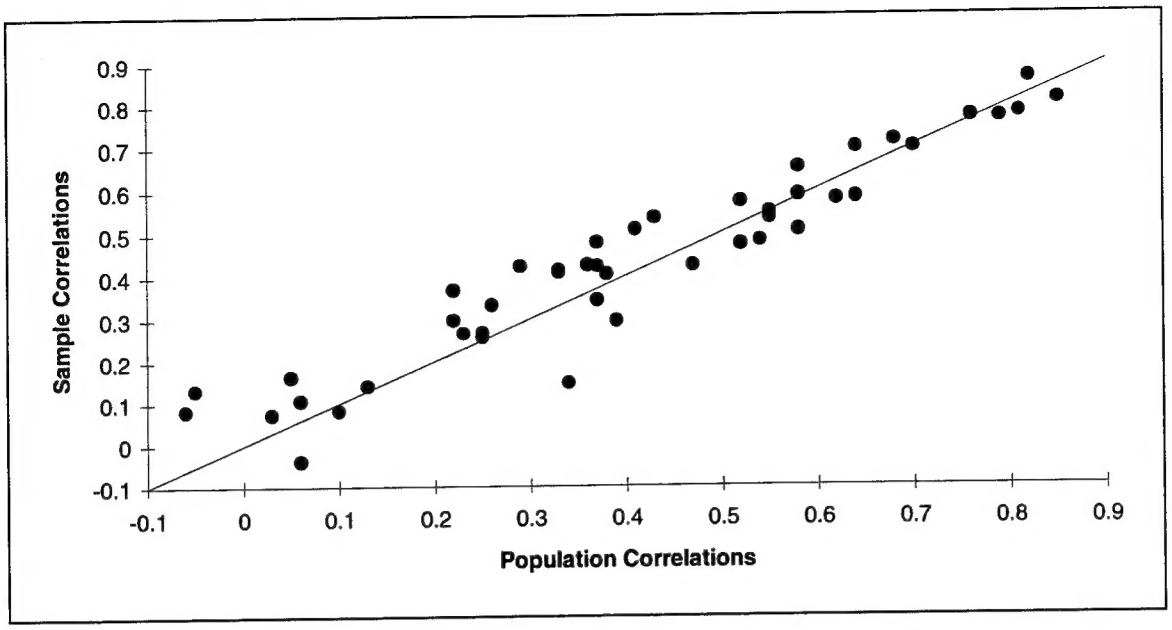

Figure 3. Comparison of Population Correlations With Sample Correlations

C. EXAMPLE 3

For this example, we generated a five-variate distribution where the first two marginals were triangular, the third and fourth marginals were beta, and the last marginal was log-normal. The desired correlations among these variables are shown in Table 4.

Table 4. Correlation Matrix for Example 3

Variable	1	2	3	4	5
1	1.00	0.73	0.64	0.40	0.15
2	0.73	1.00	0.90	0.34	0.48
3	0.64	0.90	1.00	0.29	0.42
4	0.40	0.34	0.29	1.00	0.18
5	0.15	0.48	0.42	0.18	1.00

The simulation algorithm was applied to 1,000 normal random deviates until the exact correlation matrix was attained (i.e., the *RMSE* was 0). A plot of the worst-fitting distribution (Beta Variable #1) is shown in Figure 4 and the actual and estimated moments are compared in Table 5.

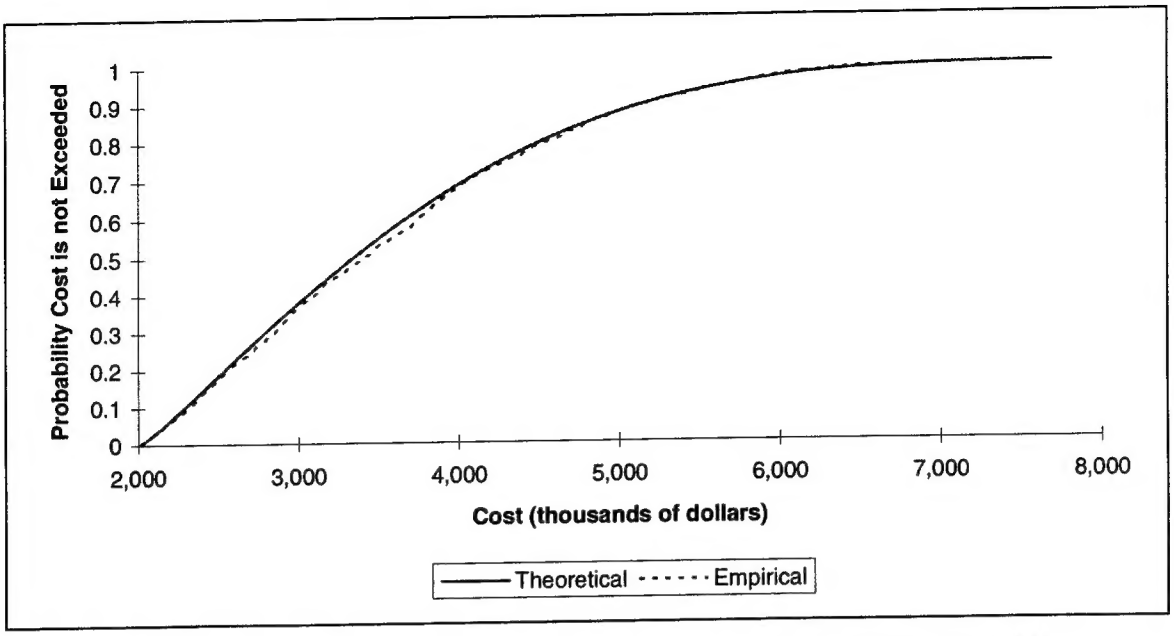

Figure 4. Theoretical and Empirical Distributions of Worst-Fitting Variable (1,000 Simulations)

Table 5. Comparison of True and Estimated Parameters From Example 3

Variable	Mean		Standard Deviation	
	True	Estimated	True	Estimated
Triangular Variable #1	2,667	2,671	514	504
Triangular Variable #2	8,000	8,017	1,472	1,466
Beta Variable #1	3,565	3,601	1,113	1,109
Beta Variable #2	9,486	9,477	1,278	1,259
Log-Normal Variable	10,000	10,010	2,000	1,953

This example shows that our simulation algorithm can be applied to a "mixed" case, where not all marginal distributions are of the same form. The computations in this example took considerably longer than in the two previous examples, approximately one hour on a VAX 4000 Model 100 computer. The increased computation arose not because of the "mixed" nature of this example, but rather because the beta and log-normal c.d.f. inversions (step 5 of the algorithm) lack closed-form solutions.

REFERENCES

[1] Johnson, Norman L., and Samuel Kotz. *Distributions in Statistics: Continuous Multivariate Distributions*. New York: John Wiley & Sons, Inc., 1972.

[2] Lee, Mei-Ling Ting. "A Family of Multivariate Density Functions With Given Marginals." Harvard Medical School, May 25, 1993.

[3] Devroye, Luc. *Non-Uniform Random Variate Generation*. New York: Springer-Verlag, 1986.

[4] Johnson, Mark E. *Multivariate Statistical Simulation*. New York: John Wiley, 1987.

[5] Parrish, Rudolph S. "Generating Random Deviates From Multivariate Pearson Distributions." *Computational Statistics and Data Analysis*, vol. 9, 1990, pp. 283-295.

[6] Fleishman, Allen I. "A Method for Simulating Non-normal Distributions." *Psychometrika*, vol. 43, no. 4, December 1978, pp. 521-532.

[7] Vale, C. David, and Vincent A. Maurelli. "Simulating Multivariate Non-Normal Distributions." *Psychometrika*, vol. 48, no. 3, September 1983, pp. 465-471.

[8] Li, Shing Ted, and Joseph L. Hammond. "Generation of Pseudo-Random Numbers with Specified Univariate Distributions and Correlation Coefficients." *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 5, 1975, pp. 557-560.

[9] Lurie, Philip M., and Matthew S. Goldberg. "Simulating Correlated Distributions With Bounded Domains." Institute for Defense Analyses, Paper P-2732, September 1992.

[10] Gill, Philip E., Walter Murray, and Margaret H. Wright. *Practical Optimization*. London: Academic Press, 1981.

[11] Marsaglia, G., and I. Olkin. "Generating Correlation Matrices." *SIAM Journal on Scientific and Statistical Computing*, vol. 5, 1984, pp. 470-475.

UNCLASSIFIED

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 1994	Final Report, Jan 1994 –Oct 1994

4. TITLE AND SUBTITLE
A Method for Simulating Correlated Random Variables From Partially Specified Distributions

5. FUNDING NUMBERS
IDA CRP 9001-707

6. AUTHOR(S)
Philip M. Lurie and Matthew S. Goldberg

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Defense Analyses
1801 N. Beauregard Street
Alexandria, VA 22311-1772

8. PERFORMING ORGANIZATION REPORT NUMBER
IDA Paper P-2998

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
FFRDC Programs
2001 N. Beauregard Street
Alexandria, VA 22311-1795

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12A. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12B. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Historical data from which to develop reliable estimates of the statistical relationships among cost elements are often scarce. Typically, information comes from many different subject-matter experts, who provide at most the moments (mean or median or mode, and possibly variance), distributional form (e.g., beta or triangular), bounds (lowest and highest possible cost), if applicable, and correlation matrix. Eliciting a fully specified multivariate distribution for the entire set of cost elements comprising a system is usually impossible. Therefore, an algorithm is needed for simulating correlated random variables from partially specified, non-normal distributions. This paper presents such an algorithm and explains how it was developed.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Risk Analyses, Simulation, Algorithms, Random Variables, Statistical Distributions	25
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	SAR

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18
298-102

UNCLASSIFIED